DOCUMENT RESUME

ED 072 099 TM 002 340

AUTHOR	Curtis, H. A., Ed.
TITLE	The Development and Management of Banks of Performance Based Test Items.
INSTITUTION	Harcourt Brace Jovanovich, Inc., New York, N.Y.
PUB DATE	Apr 72
NOTE	36p.; Symposium presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, 1972
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Compensatory Education Programs; Conference Reports; *Criterion Referenced Tests; *Information Science; *Item Banks; Migrant Education; *Performance Tests; Publishing Industry; Remedial Reading; Symposia; *Test Construction
IDENTIFIERS	Florida

ABSTRACT

Symposium papers presented at an Annual Meeting of the National Council on Measurement in Education (Chicago, 1972), all of which concern banks of test items for use in constructing criterion referenced tests, comprise this document. The first paper, "Locally Produced Item Banks" by Thomas J. Slocum, presents information on the procedures, staff requirements, and benefits when item banks are created using local staff. "Commercially Produced Item Banks: The Local Project Director's Responsibilities" by H. A. Curtis, the second paper, is based upon the author's experiences as the director of a project designed to improve the reading ability of agricultural migrant children in the elementary school of Florida. In the third paper, "Publisher's Management Problem When Entering into a New Field of Test Development" by Muriel M. Abbott, discusses the problems encountered by Harcourt Brace Jovanovich, Inc. in test development and marketing in connection with the Florida Agricultural Migrant Compensatory Reading Program. "Publisher's Role in Preparation of Items" by Barrie Wellens describes some of the unique aspects of the development of items by Harcourt Brace Jovanovich, Inc. for the Florida Agricultural Migrant Compensatory Reading Program. In the final paper, "Computer Storage and Retrieval of Test Items" by John J. Marxer, methods of item storage and retrieval are discussed, with special reference to computerized storage. (DB)

FILMED FROM BEST AVAILABLE COPY

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 072099

THE DEVELOPMENT AND MANAGEMENT OF BANKS OF PERFORMANCE BASED TEST ITEMS

H. A. Curtis, Editor
Muriel M. Abbott
John J. Marxer
Thomas J. Slocum
Barrie Wellems

Symposium presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, 1972

Published by:
Harcourt Brace Jovanovich, Inc.
New York, April 1972

The Development and Management of Banks of Performance Based Test Items[1]

H. A. Curtis, Editor
Muriel M. Abbott
John J. Marxer
Thomas J. Slocum
Barrie Wellens

Several emphases in recent educational developments have made this topic timely. The most basic of these has been the incorporation of the assessment of student learning as an integral part of the teaching-learning act. A second has been the advocacy of criterion-based instruction, with the logical necessity of assessing the attainment of each stated objective. A third emphasis has been the advocacy of accountability, which in the instructional domain logically should mean the reporting of the attainment of students in terms of reliable and valid measures of the objectives of the units with which each student has worked.

The professional literature is replete with articles advocating these ideas. The major attention has been given to the development of objectives and of programs of instruction, and relatively minor attention to meeting the measurement demands of programs of criterion based instruction or of accountability. There has been concern expressed about the technical properties of criterion based test instruments and some work on the technical properties of such instruments. A few

---

[1]Symposium presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, 1972.

federally funded centers have assembled test items from various sources, including classroom teachers and local projects generating items related to their respective projects. It probably has been the more general practice for each project to generate its own items, utilizing local personnel for the work.

In an area as new as that of making available to teachers criterion based measures of instructional programs, it is important that those who have been involved share their experiences in order that others better may understand the resources required, the costs, and the pay-offs of alternative approaches. In those settings in which local teachers are involved in writing their own objectives, and in preparing and implementing their own instructional programs, with training, the same teachers may prepare their own tests. In such situations, the costs of test production may not be separated from other costs, and important pay-offs may be the sharpening of the objectives, the production of more relevant instructional materials, and the enhancing of teacher competencies. In those projects in which objectives and programs are prepared for the use by schools outside the specific project, the personnel required for test production may not be available, and the cost of test production must, or should, be a distinct budgetary item. Furthermore, the payoffs may not be applicable.

In the sections which follow, Slocum's report is based upon his experiences in two local settings, the Downers Grove, Illinois, Public School System and the Center for Educational Development of the University of Illinois College of Medicine. Curtis, Abbott and Wellens report from their several vantage points their experiences in a project in which a

test publisher prepared test items to measure previously determined objectives. Marxer reports on the basic requirements of storage and retrieval systems designed to make stored items available for future use.

Locally Produced Item Banks

Thomas J. Slocum
Center for Educational Development
University of Illinois College of Medicine

The purpose of this paper is to present information on the procedures, staff requirements, and benefits when item banks are created using local staff.

Procedures

1. Specification of Learning Outcomes

To provide clear directives for evaluation, specifications of learning outcomes should be stated. Each statement should include the specific content and observable behavior to be mastered by students as well as the conditions under which such mastery will be demonstrated.

2. Division of Item Construction Responsibilities

Responsibility for producing test items may be distributed among the staff who are to prepare the items.

3. Review and Editing of Items

Each item should be checked to determine that it indeed poses a problem for the examinee that would require him to demonstrate mastery of the learning outcome for which the item was developed.

Each item should be checked for technical soundness. Flaws in construction can include grammatical errors, unclear instructions, and the like.

4. Coding for Retrieval

Following completion of editing and revision, each item should be coded for retrieval. Possible classifications include the learning outcome, content, behavior, source or reference, course or class level appropriateness. The item should be coded on any characteristic that might be the object of a request for items.

Such coding information can be keypunched using computer cards. Sorting of the cards would then provide lists of items in each classification or combination of classifications. If each item is assigned a unique identification number then this number can be the output of the retrieval system. Such numbers can be listed under any of the categories used in the coding stage.

5. Retrieval

The possible ways of storing coding information ranges from printing lists of item information in a form similar to the white and yellow pages of the telephone directory to computer storage of the coding information and the items. That is for each code term the identification numbers of relevant items are listed.

One way of storing test items is to keypunch the items. Each line of text of the item is punched on a card. The set of cards for an item can be listed on continuous ditto-master. Another way to store items is of course as they are written. Each time an item is to be used it would then be retyped. At least one school district pastes a printed copy of an item on the coding card described above.

6. Pilot Testing of Items

The initial administration of an item may result in the identification of errors in item construction. Items that prove to be unsatisfactory should be sent back to the review and edit stage and re-routed.

7. Administration and Record Keeping

As an item is used the resulting data should be stored with the item so that the item can be improved if needed and to aid in the selection of items from the pool.

Staff Requirements

The procedures listed above indicate the need for a director either to conduct inservice training in the specification of learning outcomes and construction and editing of test items or to arrange for outside help with the inservice training and other procedures. In addition, responsibility for item construction and editing responsibilities would have to be divided appropriately and fairly. Furthermore, all efforts would have to be scheduled.

It seems that no attempt to produce an item bank locally could succeed without the interest and support of the local faculty members. Release from other duties or support during vacation time would have to be provided for faculty members to be available for item construction. Two successful local item development arrangements are described in following paragraphs to illustrate how these requirements can be met.

Local Projects

The first project entitled the "Evaluation for Individualized Instruction" (EII), was supported by a three-year grant to the Downers

Grove, Illinois school district. The grant was administered by the Institute for Educational Research which is located in the same city. Classroom teachers were employed for three to nine-week periods to write behavioral objectives and objective test items. Staff members of the Evaluation Project conducted appropriate inservice training of the teachers and created a computerized retrieval system. Approximately 5000 behavioral objectives and 25,000 objective test items are now available.

The second is an on-going project conducted by the Center for Educational Development (CED) at the University of Illinois College of Medicine. CED employs evaluation specialists to work with medical school department representatives who are responsible for the preparation of comprehensive examinations for each of the first two years of medical school and the two clinical years. One evaluator works with each committee.

The faculty members who are directly involved with the construction of the examinations obtain items from their departmental colleagues after general guidelines for the content and behavior to be measured are outlined. The CED representative assists the faculty in the production, review and editing of the items. Several hundred new items are generated each year. Approximately 7,000 items from previous comprehensive examinations are on file for faculty use. About 3,000 other items are available to the students for self-study. The self-study items are presented using a computer and cathode ray tube-keyboard terminal. The comprehensive items are coded by content and behavior as well as department of origin and statistical results. This coding information is stored in a computer-based retrieval system.

From each of these projects a number of benefits were apparent. The availability of items for formative evaluation, individualized instruction, or criterion-based instruction may be the greatest benefit derived from local item development, especially since tests that measure student attainment of objectives specified by local staff can be made easily.

The effect on the faculty members who receive training and help in formulating their objectives for their students in behavioral terms has been to aid the teacher in communicating with other teachers and with students. The feedback to students and teachers can be more precise and more timely. Duplication of effort in producing test items can be reduced to the point of practical elimination.

Conclusions

The writer has observed the operation of each of the two projects reported. He recognizes the importance of the benefits cited above. His knowledge of the resources required and of the costs involved in local item production have caused him to consider carefully whether substantially the same benefits might be realized if the actual production of items were not an integral part of the local faculty's responsibilities. It is his present conclusion that if local faculty members are trained in using behavioral specifications of learning outcomes and in recognizing whether an item would measure the presence of a desired outcome, then many of the advantages of local item projects could be obtained from the use of existing banks of objectives (learning outcomes) and of test items such as the Evaluation for Individualized Instruction Project, Downers Grove, Illinois, or the Instruction Objectives Exchange, UCLA Graduate School of

Education, Los Angeles, California.

This may be a startling conclusion. In light of the fact that one project had to discard about one-half of the objectives and items generated by its teacher-participants, others who would start local item projects might best begin by using existing banks or pools of items to provide local staff with albeit vicarious, but important, experience in using behavioral learning specifications and objective measures of higher mental processes.

Commercially Produced Item Banks: The Local Project Director's Responsibilities

H. A. Curtis
Florida State University

This report is based upon the writer's experiences as the director of a project designed to improve the reading ability of agricultural migrant children in the elementary schools of the State of Florida. A catalog of objectives had been prepared by the project staff and it had the responsibility of supplying more than 2500 items for the measurement of the objectives. The project was not responsible for the preparation of the instructional program, or programs, nor for the training of teachers who were expected to carry on the instruction. Furthermore, the project did not have at its command the services of a sizable group of teachers experienced in such instruction. Thus the writer's responsibilities, and the resources at his command, differed materially from those of the director of a project designed to serve a specific local instructional program. For these reasons, the consideration of alternative approaches to item production seemed reasonable.

There has been a test publishing industry for fifty years. Educators have learned to turn to this industry for psychological tests measuring both broad and specific abilities, and achievement tests in broad subject areas. As a consequence, the several major publishers have developed the capability of supplying published tests, and each has its staff of skilled item writers and editors. The writer reasoned that the test publishers

did have the capability to produce items, which is precisely the capability lacked by many local schools and R & D projects, his own in particular.

The recruiting, training and supervising of item writers was considered. The time and costs were weighed in terms of the probable productivity of such trainees, and the decision to turn to the publishing industry was made.

At the outset, it became clear that we all were at a strange impasse. The first fact that was established is that while sophisticated publishers have relatively large stocks of unused test items, many of such items were not readily available for use in this project. The reasons are to be found in the process of test publication. As an illustration, the publisher may obtain 150 or more items from its item writers on a royalty or fee basis when it expects its finished product to require only 60 items. All are first scanned and unsuitable items laid aside, unedited. Perhaps 100 are edited, and subjected to tryout. The tryout data may indicate poor measurement qualities for some, and they are laid aside without further work. Of the surviving items, there likely will be surpluses in some categories but not others. Surplus items are laid aside, and the remainder subjected to further editorial scrutiny and finally included in the published test. For these reasons, and for the further reason that test items become "dated," the unused test items in publishers' files did not constitute a readily available pool from which immediately usable items could be drawn.

The second fact that became evident was that while the need which this project presented had long been predicted, no provision had been made to meet it. Those who have been in measurement for any length of time have foreseen the need for banks of items from which withdrawals could be made

to develop instruments to measure the specific outcomes of instruction and the results of specialized and experimental programs. Each of the major publishers has recognized this need too, perhaps even more clearly than have local educators and measurement specialists. One publisher commented that there was a memo in its files, dated 1938, warning that this movement was coming, and urging the company to prepare for it!

A third fact that became apparent was that the writer, as a representative of the measurement and research fraternity, and the members of the professional staff of the publishing industry really did not know how to talk to each other. Experience in requesting items from the publisher's hypothetical banks is not a part of our lore, and neither is responding to such requests a part of the publisher's lore.

When it dawned on all of us that publishers really had no available item pools, that the future that we had all seen coming was here now, and that we were all inexperienced in working on such an undertaking, the air somehow cleared, and we turned to the question of what each party could and should contribute to the joint undertaking.

Speaking now as the project director (Drs. Abbott and Wellens will speak for the publisher), the first input that must be made by the project director is a clear description of the program and of the population to be served by the program for which test items are desired. This input serves three purposes. First, it gives the professional staff of the publisher a feeling for the subjects and for the situations to be served by their end product. The making of this input should not be hurried, because during the presentation, each of the members of the publisher's staff searches his

own background for experiences and products that may be relevant to the situation being presented. Obviously, the more of these that each can recall, the more the publisher will have to start to work with. Second, the description of the subjects and of the settings of their lives gives the professional staff members extremely valuable clues to the topical bases, interest areas, and literary qualities that will appeal to the subjects and thus contribute to the appeal of the items to be produced. Third, the description of the program, of its manner of operation, limitations and duration supplies the substantive basis for making decisions about the terms of the business arrangements that should be made. These business arrangements include provisions for item security, specification of the extent of usage, the basis of procurement (purchase or lease), the time period to be covered, the official contracting agencies, and the identification of the parties to the agreements that must be finalized.

The second input that the project director must make is the supplying of his catalog of objectives. It is obvious that this catalog serves as the table of specifications of the desired test items. It is equally obvious that supplying this catalog is the project director's responsibility. Three things may not be so obvious. First, among the objectives stated there may be those which are not directly testable, but require observational techniques on the part of teachers actually presenting the program, indirect and "unobtrusive" measurement, anecdotal treatment, etc. These should be set aside as objectives for which the publisher will have no responsibility, and the decision to set them aside can be made most constructively in conference with the publisher. A second point that must not

be overlooked is that the project director's objectives were prepared by a local staff whose statements of objectives carried precise denotative meanings to them, but not always to the outside item producer. The director should be prepared to go over the catalog, objective by objective, with the publisher's staff to make certain that the denotative meaning of each objective is clear. A third point is that for certain objectives, the contexts within which measurement is desired should be specified by the project director. For instance, certain critical reading objectives may be measured in the context of propaganda hand-outs, of mail-order catalogs, or of formal philosophical arguments. Better items will result if the director makes clear his choice of the contexts within which the items should be prepared.

The third input which seems to this speaker to be the project director's responsibility is the supplying of illustrative instructional materials, particularly materials used in specialized programs or programs designed for a deviant subset of the general population. To illustrate, item writers generally are familiar with standard textbook and library materials. Also fairly well known are materials prepared for use by disadvantaged children in urban areas, and newspapers and incidental reading materials circulating in urban areas. But project directors whose target population is the rural segment of the poverty stratum should be prepared to furnish small farming town newspapers, small church notices, small town handbills, etc. Simply because only the project director can supply some of such materials, and because in other cases he can do so more efficiently than can anyone else, he should be prepared to take the responsibility of supplying illustrative materials as needed.

Representing the local project in all contractual matters is a central responsibility of the project director. In our experience, the formulation of the contract went through two identifiable stages. The first I shall call the Educator's Draft, and the second, the Attorney's Draft. The Educator's Draft was a document that identified the items to be developed, the financial terms, and the duration and conditions under which the items could be used, all in a language which was perfectly clear to the publisher's staff and to the project director. Clearly, the project director should represent the project in the preparation of this draft to make certain that the local project gets what it needs and under conditions favorable to its use.

The role of the project director in the formulation of the second draft is quite a different matter. Under the best of conditions, the members of the director's local legal staff may be assumed to be experts in their field, but the assumption that they are able to read the educator's language with sufficient understanding to translate it into contract language may not be tenable. Suffice it to say that the project director should be prepared to work patiently with his legal staff to develop basic understandings of the essential elements of the contract so his attorneys can draft correctly the proper legal document.

The publisher's representative probably is going through the same process with his legal staff at the same time. Close liaison should be maintained between the educational professionals to minimize the final adjustments that in the end will be made by two sets of attorneys.

While it may not be true in all cases, in our case the project director also had to work with his own purchasing and disbursing departments to be certain that the developing contracts and other documents

would be in such form and in such order that the publisher would actually be paid when the job was done.

Finally, the project director should read critically each and every item before its final approval. If items are in the process of production, the reading of each item in an editorial manner is productive. The project director can compare the sense of the item with the sense of the objective, detect regional and cultural biases that will affect adversely the validity of an item when administered to his group, and contribute to the elimination of just plain "bugs" in the items. If it is a matter of item selection, the reading is necessary to establish the relevance of each item to its objective, suitability for the local population, and its administrative feasibility under local conditions.

In conclusion, the writer wishes to offer as his considered judgment the statement that test publishers can and will deliver a most valuable service if the project director is prepared to discharge his responsibilities, and if cooperative and supportive relationships are established at the beginning and maintained throughout the negotiations, and throughout the development and/or selection of the items.

Publisher's Management Problems: When Entering into a New Field of Test Development

Muriel M. Abbott
Harcourt Brace Jovanovich, Inc.

Whenever a test publisher or test development agency enters into a new area of test development or marketing, it is faced with unfamiliar problems that require new approaches and new solutions. Involvement in the Florida Agricultural Migrant Compensatory Reading Program (Florida) provided an opportunity for Harcourt Brace Jovanovich, Inc. (Harcourt) to participate in a new approach to test development and marketing. It also presented an opportunity to use the product resulting from that experience as a basis for the development of a new assessment service to educators. The problems that arose from this undertaking concerned not only those of a professional nature in the area of item development, but also those of a very practical and legal nature in the area of producing and marketing items, and in ensuring the protection of both the consumer and the producer of items under various conditions of item use. The solutions to many of these problems were possible only through the flexibility and close cooperation of the staffs of the two participating organizations.

The Florida group basically was using a systems approach in its reading program for migrant pupils. They had established a set of behavioral objectives, had assembled programs designed to lead to the achievement of the objectives, had planned for the development of a system of measurement, and had, also in the planning stage, a method of reporting and evaluating results. The Harcourt Test Department was

approached to determine whether or not it would be interested in participating in the measurement phase of this program. Harcourt's responsibility in this project would be the development of items designed to measure the specific reading objectives.

This undertaking had immediate appeal. It not only offered an opportunity to participate actively in a systems approach to an educational problem but also an opportunity to develop and implement a more flexible assessment system than the more typical test publishers' undertaking which is to provide a predetermined instrument for assessment purposes. This opportunity arose from the fact that Florida was willing to consider a lease agreement rather than outright purchase of items. The lease agreement provided to Florida a cost advantage and to Harcourt the opportunity to use the items developed to implement its item bank. The item bank had long been considered as an efficient way to provide assessment instruments tailored to meet specific consumer needs. Because the items to be developed for Florida were designed to measure specific objectives and not a particular program, they were eminently suitable for this purpose.

The development of items to measure objectives without any reference to program or curriculum material by the item developers presented an interesting situation. However, if a program is designed to lead to the achievement of specified objectives and items are designed to measure achievement of these same objectives, then both the learning program and the items can be developed independently. This has not often been done in practice. The more usual procedure has been to relate measurement to program or curriculum rather than to relate measurement and program through the same set of objectives.

The advantages and appeal of the undertaking were clear from the beginning as was the knowledge that many unfamiliar problems would arise. The first problem to be approached concerned item source, or where to obtain, within only a few months, over 2500 items designed to fit detailed specifications. Consideration was given to the use of items from existing pools, such as items constructed and tried out in the process of developing a number of Harcourt published tests. Harcourt, however, had reservations with respect to their use. Scanning hundreds of items and attempting to assign them to an appropriate objective is not efficient. It is time consuming and apt to result in incorrect assignment. Also, not only would some of these items be "dated" or inappropriate in content, but inevitably, there would be "gaps"; objectives with no items to measure them. Legal problems in connection with using available items on hand arose in connection with authorial royalty agreements. It would have been a considerable undertaking to determine how royalties should be paid on a mix of items consisting of those to which all rights are held solely by Harcourt and items derived from different published instruments to which authors also have certain rights. These instruments, moreover, have not only different authors but also different authorial agreements.

The Florida staff appreciated Harcourt's position and, although acknowledging that items with data were preferable, agreed that the sole source of items would be those developed specifically to measure the identified behavioral objectives. It then became necessary to set up an organizational system that would ensure the obtaining of the required items within the limited time period. Of primary importance was the securing of a sufficient number of experienced and competent item writers

as well as editors who would work within the time and item specification constraints. The task of developing items that would measure a particular behavioral objective, yet be uniquely different from items measuring a closely related or similar objective, was somewhat different from that typically encountered in Harcourt test development. Therefore, considerable training and supervision of writers and editors was required. The importance of competent and extensive editing of items by trained editors cannot be overemphasized. Because of the tremendous amount of item writing, editing, and rewriting that was done, instituting and carrying out procedures to control the flow of more than 2500 items within the tight time schedule was no small task in itself.

Another problem concerned the obtaining of item performance data. These data were obviously desirable to both organizations. Because of time constraints no tryout of items was possible and any such data had to be obtained from the assessment program itself. A very real concern was what derivable data could be meaningfully interpreted. Another major consideration was the fact that although Florida would use the items in a criterion-referenced instrument, potential users of the items might want to include them in an instrument for criterion and/or normative interpretation.

Consideration was given to the fact that in the Florida project the data would be derived from a group atypical of the United States school population as a whole. The Florida sample was describable, however, and the data obtained could be interpreted accordingly. Furthermore, with respect to items in the item bank, there are definite advantages in accumulating item performance data derived from clearly identified but different educational programs and types of children.

Item difficulty and data reflecting the attractiveness of each mislead are of value whether items are to be included in a norm or criterion-referenced instrument. Because, for any particular item, difficulty is relatively independent of total test score or of the item mix in which it is included, the interpretation of these data, when derived, would present no problem. In the case of item discrimination, however, a different situation exists as these data do depend upon total test score which in turn is related to the particular set of items. Because items in the item bank, as well as in the Florida program, will not be assigned to an invariant set but may be used in any number of different combination of items, it was recognized that meaningful interpretation of item discrimination data was limited.

The introduction of a different marketing mode gave rise to unfamiliar problems in estimating costs and price to consumers. Major problems concerned the nature of the marketing unit and the conditions under which the unit was to be marketed. Traditionally, the marketing unit has usually been a test consisting of a copyrighted set of items. Under the new approach, the unit is a copyrighted item. When a test is the unit, developmental cost and price to consumer is determined for the group of items as a whole. It is not necessary to determine the specific cost of each individual item. When an item is the marketing unit, however, a different situation exists. If consumers are to be offered items on an unrestricted selection basis, developmental cost and price to consumer must be determined for each particular item. Costs for different items or item types vary greatly; for example, a reading passage, together with an item based upon it, has a cost many times that of a vocabulary synonym item.

Product lease rather than sale gave rise to the other major marketing problem. The conditions under which a test publisher has usually marketed its products has been to retain copyright ownership and proprietary rights to a product and sell to consumers printed ready-for-use copies of the copyrighted product or test. In this situation, for a particular test, conditions of sale are the same for each consumer and difference in cost is based solely upon extent of use or number of copies purchased. For Florida, however, a leasing arrangement was introduced. Under a leasing arrangement the producer retains copyright ownership and proprietary rights to that product but leases to the consumer the right to use, print, publish, and reproduce the product subject to contractual stipulations. In this case a single copy of the product, for example, an item, is delivered to the consumer who then determines the form in which the product is to be used. Conditions of lease can vary with each consumer and cost varies with the conditions. Variable conditions include duration of lease, restriction on use, extent of use, etc. It should be noted that for a consumer the leasing of items offers a considerable price advantage over their sale, as no single user then bears the entire developmental cost. To date, the problem of precise determination of fees under the different conditions of lease has not yet been fully resolved.

Legal problems arose in the areas of protecting the copyright and proprietary rights of the producer while protecting item security for the consumer. Whenever copyrightable materials are published, appropriate copyright notices must be affixed. This is relatively simple in the case of a predetermined printed set of items. However, when items are supplied in individual form, a very different situation exists. Merely sending any

uncopyrighted material through the mail can be construed as publication and, therefore, copyright protection is necessary or the materials will be in the public domain. Because each item to be sent was printed on an individual sheet, the appropriate notice had to be affixed to each item. Also, copyright provision had to be made for the future situation when the items would be reproduced in sets. Furthermore, Florida had indicated that there was a possibility that they might want to include Harcourt items in a set with items from other sources. These other items might or might not be copyrighted by other agencies. This was a unique situation not only for the Harcourt Test Department, but also for its lawyers. Since it was obviously impractical to print a Harcourt copyright notice beside each Harcourt item appearing in a set, a practical solution was finally achieved that provided for a general copyright notice to accompany the set.

Florida needed assurance of item security both with respect to a sufficient time period of lease to ensure program implementation and with respect to protection of the item from exposure to potential examinees. Harcourt wanted maximum freedom to lease the items to other customers. Different methods of ensuring item security were considered. The method adopted by Florida guaranteed geographical restriction of use. This agreement provided that for a specified period of time the use of items in the State of Florida was restricted to the Migrant Compensatory Reading Program with the use of the items prohibited to any other program, person, or entity within that State. Outside the State of Florida, however, Harcourt retained exclusive rights to lease any or all of the items to any agency. The lease is subject to renegotiation upon its expiration.

It should be noted that the type of contract that was finally drawn required very different provisions from the customary Harcourt agreements. A single agreement had to ensure both protection of copyright to the producer and protection of item use to the consumer under all possible conditions of item administration while, at the same time, ensuring the availability and protection of these items to potential consumers. Working out these provisions required a continuous exchange of information and suggestions between the professional staffs of the two organizations and their attorneys as the legal drafting of the contract proceeded. It was a protracted and painstaking process but finally resulted in a document that will serve as a guide in similar future transactions.

The successful completion of the Florida-Harcourt project resulted in no small measure from the close cooperation of the staffs of the two organizations and their sympathetic understanding of each other's problems. The importance of this kind of working relationship cannot be overestimated. It is crucial if the supplier of an assessment service is to tailor his product to meet the specific assessment needs of a particular educational organization. Indeed, the staff of Harcourt found that its experience working closely with the staff of an educational organization to investigate a problem, explore possible avenues of approach, and arrive at mutually satisfactory solutions was most rewarding. This type of cooperative venture or "new alliance" between test publishers and educational organizations offers great potential for the solution of other and even more complex problems confronting education today.

Publisher's Role in Preparation of Items

Barrie Wellens
Harcourt Brace Jovanovich, Inc.

This paper will describe some of the unique aspects of the development of items for the Florida Agricultural Migrant Compensatory Reading Program (Florida). The publisher's task was to develop test items which measure attainment of Florida's reading objectives in the most efficient way possible. By "efficient," we mean that the item fully measures the objective, testing time is used effectively, administration and scoring are as simple as possible, and expense is kept to a minimum.

The publisher, Harcourt Brace Jovanovich, Inc., (Harcourt), was given Florida's 162 reading objectives and asked to devise ways of measuring each at the grade levels specified. While every effort was made to work with each objective as it was stated, when changes were thought to be necessary, the Florida staff worked with the publisher to modify the objective so that it could be measured more effectively. In a few cases, two objectives were so similar that it was wondered whether the distinction had any real meaning. With Florida's help, a different way of measuring each was devised, but it will not be known until the items are administered whether or not a difference actually does exist.

Many of the objectives could not be measured by group administered paper and pencil items, for example, "The learner will demonstrate the ability to organize his thoughts and to present them orally in a logical manner." Fortunately, Florida's testing facilities permit great flexibility with respect to mode of presentation of items, mode of response,

and scoring. Each objective could be measured as Harcourt and Florida thought it should be measured. It was not necessary to force items into the multiple-choice mold where it was inappropriate.

Items could be either group or individually administered. Items or parts of items could be dictated on tape, printed, or projected on a screen in the form of slides or film strips. Responses could be either oral or written, and could range from multiple-choice to open-ended to task performance. The scoring guide accompanying each item would be a key for multiple-choice or arrangement items, and a list of acceptable and unacceptable responses or a set of criteria for free response items. Scoring guides could also be in the form of a taped standard against which the examiner judges the pupil's performance.

Certain guidelines were followed in deciding how to measure the objectives. Mode of presentation had to be as simple as possible, especially at the lower grade levels. Mode of response had to be direct. For example, where the task is to unscramble a set of pictures presented in random order, rather than having the pupil choose the proper order from four options, he is asked simply to number the pictures in order. This also serves to increase the number of options from four to twenty-four when four pictures are used. (Of course, this can be easily converted into a multiple-choice item if desired.) The major guideline was that paper and pencil group administered items were to be used wherever possible.

Harcourt, in collaboration with Florida, devised ways to measure each objective at each designated grade level. This resulted in fifty-five basic structures or formats which we call item types. Most item

types can be used in many situations. For example, the item type "80-100-word passage + task + scoring guide -- oral response" was used for 104 items measuring ten different objectives.

Once basic methods of measurement had been resolved, a prototype* was written for each objective. Whenever one of the 162 objectives was to be measured at more than one grade level, a different prototype was written for each grade level, resulting in a total of 269 prototypes or sets of specifications. The prototype is the item type applied to a particular objective at a particular grade level. Many prototypes were generated from any one item type. A prototype includes the statement of the objective and its purpose; the grade level; the number of items required; the item type designation; estimated administration time; descriptions of the stimulus, mode of response, task, and scoring guide; and one or more examples from which to generate items. (For an illustration of an item type and a prototype, see Attachment.)

A variety of approaches was especially important where more than one objective dealt with the same general ability. Therefore, wherever possible, a prototype contained more than one sample item to show different ways of measuring the objective. For example, the ability to carry out written directions can be measured by an item in which the pupil is given a picture of two rows of assorted shapes along with directions such as "Put all the circles in the top row. Below each circle, put a square." He is then given four pictures and asked to choose the one showing that the directions were followed. The same ability can be measured by having the pupil follow written directions for changing the batteries in a

*It should be noted that what National Assessment of Educational Progress calls a prototype, Harcourt Brace Jovanovich calls an item type.

flashlight using actual equipment. (In this case, he gets immediate reinforcement: If it lights up, he knows he's right.)

Now, 2500 items had to be written from the prototypes. Nineteen item writers participated in the project. Although they were experienced, the item writers had to be tried out on different item types in different areas of reading so that appropriate prototypes could be assigned to each individual.

In view of Harcourt's dual purpose in preparing the items, it was important that item content be suitable for pupils throughout the country as well as for the pupils in the migrant program who come from a distinct cultural group. For the most part, this meant avoiding content which was inappropriate for the migrant children. Content specifically relevant to their daily lives was included in a separate section called "Applications." Dr. Curtis met with the Harcourt staff in order to describe in detail the life style of the migrant population and the educational problems confronting the children, and this information was passed on to the item writers along with source materials which had been gathered by Florida personnel.

Among the source materials used for the Applications section were local Florida newspapers and pamphlets from the Florida Department of Health and the Florida Institute of Food and Agricultural Sciences on topics such as home economics, health and family, and safety. Social Security publications and mail order catalogs were also used.

Items were to be written to assess reading at a level suitable for the typical pupil in the United States in Grades 3 through 6 even though the migrant pupil tends to be older than most pupils at the grade level at

which he is functioning.

The next phase, the writing, editing, and rewriting of items, involved interaction between Harcourt editors and the item writers, the artist, and the Florida staff. Since, at any given time, each item was in a different stage of development, keeping track of the 2500 items was a job in itself.

Each set of items written from a prototype was submitted to at least three editors working independently; the pooled judgments were then incorporated into the final item. One of our biggest problems was obtaining enough good, interesting reading passages. Another problem was that items did not always measure the objectives for which they were written. For items that we knew were going to be difficult to write, the item writer was asked to submit one or two samples before writing all the items required. This saved much time and effort.

Even though more than one example was given in the specifications for some objectives, it was found that some item writers chose to measure the objective one way only. For example, "The pupil will identify at least one rational reason why a certain statement has, or has not, been proved in a given passage" can be interpreted in three ways: "The statement was proved because;" "The statement was not proved because:" or a combination of the two. The item writer found the first way to be the easiest and neglected the others. The items were returned to the item writer for correction, but time was lost in the process.

Once the method of measuring an objective had been decided, the most crucial phase in item development proved to be the editing. Never have we been so impressed by the importance of thorough, professional editing.

While creative item writers do furnish many ideas, it is editing that makes or breaks the items, particularly under severe time pressure.

Looking back on the total project, it was certainly a challenge to be presented with a new measurement task -- writing items to measure specified objectives instead of general areas or established programs. We analyzed objectives as we have never analyzed them before. We discovered that we could use many modes of presentation and of response beyond those generally used in published tests. We are firmly convinced that the critical importance of editing can never be overestimated. And, finally, we learned through experience that the prime factor contributing to the success of any undertaking of this nature is the close cooperation between the two organizations involved. The free exchange of thoughts and mutual resolution of difficulties contributed immeasurably to the development of a better product.

Publisher's Role in Preparation of Items

ATTACHMENT

ILLUSTRATION OF ITEM TYPE AND PROTOTYPE

Item Type

Item Type CT = "Components of critical thinking" passage + multiple-choice item, 4 options.

Prototype

Objective No. IV-19 Item Type: CT Grade: 6.8

Major Category:	Comprehension
Subcategory:	C. Critical Reading 1. Logic
Objective:	The learner will be able to identify illogical thinking, inconsistencies, fallacies or discrepancies in a given selection.
Stimulus:	Brief passage (25-50 words) or syllogism and one multiple-choice item
Number required:	8
Estimated Administration Time:	2 minutes per item
Mode of Presentation:	Printed on a page
Mode of Response:	Multiple-choice written
Scoring Guide:	Keyed response

(Continued)

Examples: A. Rose and Brenda were having an argument. Rose said, "There are 48 states in the United States. I remember learning that last year." Brenda said, "You're wrong. Our teacher told us yesterday that there are 50 states." Finally, Rose said, "I know I'm right because I'm older than you are."

What is wrong with Rose's reasoning?

a. Brenda may have misunderstood her teacher.

b. Rose is talking from memory, not fact.

c. Age has nothing to do with being right or wrong.

d. There is no right or wrong; it's a matter of opinion.

Key: c

B. 1. I collected just as many stones as Mark.
2. Dan and Steve each has as many stones as Mark.
3. So I guess I have more stones than Steve.

In order for sentence 3 to be correct, it should say --

a. I have as many stones as Steve.
b. I have more stones than Dan.
c. Mark has more stones than Steve.
d. Dan has the most stones.

Key: a

C. All the girls in my class live on Main Street or Broad Avenue. Most of the girls have older brothers.

Which conclusion is false?

a. Some of the families on Broad Avenue have at least two children.

b. Some boys on Main Street have younger sisters.

c. Mary Ellen, who lives on River Road, is in my class.

d. Joan is in my class, so she must live on Broad Avenue or Main Street.

Key: c

Computer Storage and Retrieval of Test Items

John J. Marxer
Center for Educational Development
University of Illinois
College of Medicine

The other papers in this series have focused on methods of producing items. We shall concentrate on methods of item storage and retrieval with special reference to computerized storage. Our experience in computer storage of items indicates that this is a profitable venture in terms both of usage of the items and of keeping the items updated.

The Computer Systems Section of the Center for Educational Development (CED) has produced a system known as CRIB, a Computerized Random Item Bank. This is an on-line, real-time system in which students use self-selected items on computer terminals and receive immediate feedback of their results. The system contains about 2500 items that are categorized by discipline and subdiscipline. For example, Anatomy is a discipline and Morphology is a subdiscipline within Anatomy.

Students use the system on any of 10 available terminals simply by typing in their registration number and then selecting the area in which they wish to be tested. The system tells them after each item whether their choice was correct and also keeps a cumulative score for each student to which he can refer whenever he pleases. The system is designed for self-evaluation so that an individual's scores are not available to the faculty, but results for individual items or groups of items are available to faculty members.

CRIB is written in Coursewriter III, an IBM Corporation language, and

runs on an IBM 370 Model 155 computer. The creation of our system was basically very simple. Items are stored in the same form they are presented to the students and they are divided by area into different sequences of labels. Once the student has chosen an area, a random number generator is used to select items from that area and present each item to the student. All items on our system are multiple choice items with 9 or fewer choices, but this is a system limitation that could be changed without too much difficulty.

The items stored in CRIB were obtained from files of past comprehensive examinations at the University of Illinois College of Medicine. Thus we did not face the task of actually constructing items. We did categorize the items by discipline and subdiscipline and this task was done for us by the Evaluation Section of CED and faculty members in the departments that originally wrote the items.

Items were and are keyed into CRIB by a secretary. They are then checked for accuracy and the central coding modified so that they are available to students. We ask students using CRIB to comment on items they feel are incorrect or outdated. Outdated items seem to occur with some frequency in the medical sciences where new methods of diagnosis and treatment become available every month. Students evidence a good deal of enthusiasm for CRIB and their notes regarding item changes have been a great help to us in keeping the items in the bank updated.

Our system obviously has a rather rigid selection algorithm based on a predetermined categorization of each item. Another type of item bank exemplified by the one in use at Wayne State Medical School has items

(or item abstracts) stored with attached keywords and/or various parameters such as item statistics. Their system is not on-line to students but rather a batch system designed to select groups of items from the bank based on requests which may contain any logical combination(s) of keyword(s) and/or other parameter(s).

Essentially then there are several possible types of systems of item storage. Systems can be on-line to students or batch systems. Items can be stored with fixed categories or with attached descriptors. Fixed categories require less computer time to store and to retrieve items but offer less flexibility in retrieval and may become obsolete as curricula change. They also tend to limit the possibilities for the exchange of items between institutions. Descriptors attached to items offer more flexible retrieval possibilities but require more input and processing time.

In conclusion both types of item banks mentioned seem to have unique capabilities for storage and retrieval of items. We urge the creation of more such banks in order that the possibilities they present may be thoroughly explored. We recommend that those who begin to build such item banks communicate with each other and those who already have such systems in an effort to develop compatible methods of storage so that items may be freely exchanged between systems for the benefit of all.